Network Simulation Models

Rolf Riesen
Sandia National Laboratories
rolf@sandia.gov

July 23, 2009

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

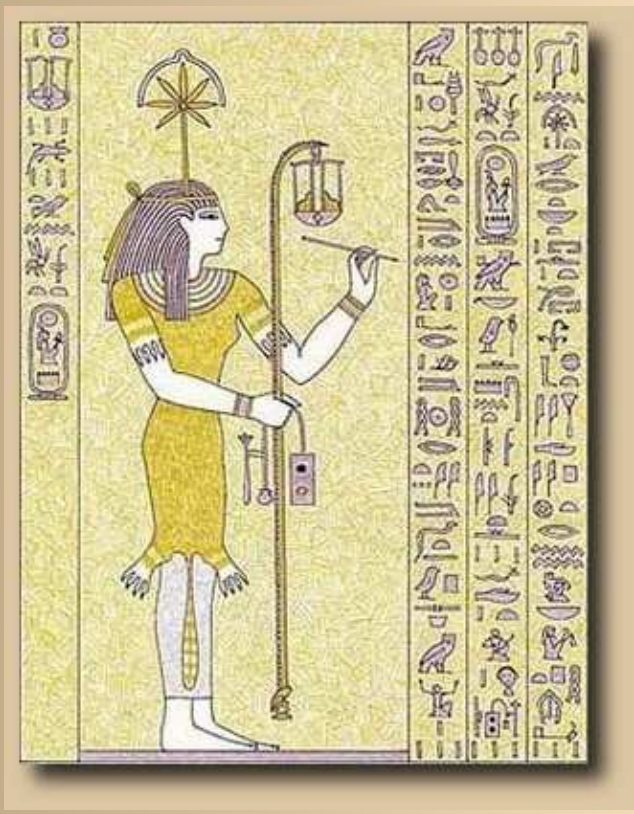
Motivation

- Simulating a large, next-generation system to the gate level is not always possible, nor necessary
- Sometimes a less detailed approach is enough

Seshat

- Named after Egyptian goddess responsible for measurements and record keeping
- Execution driven network model
- Inserts between application and MPI library
- Runs application in virtual time frame
  - Model dictates delivery of actual data
- Models XT-3 Red Storm

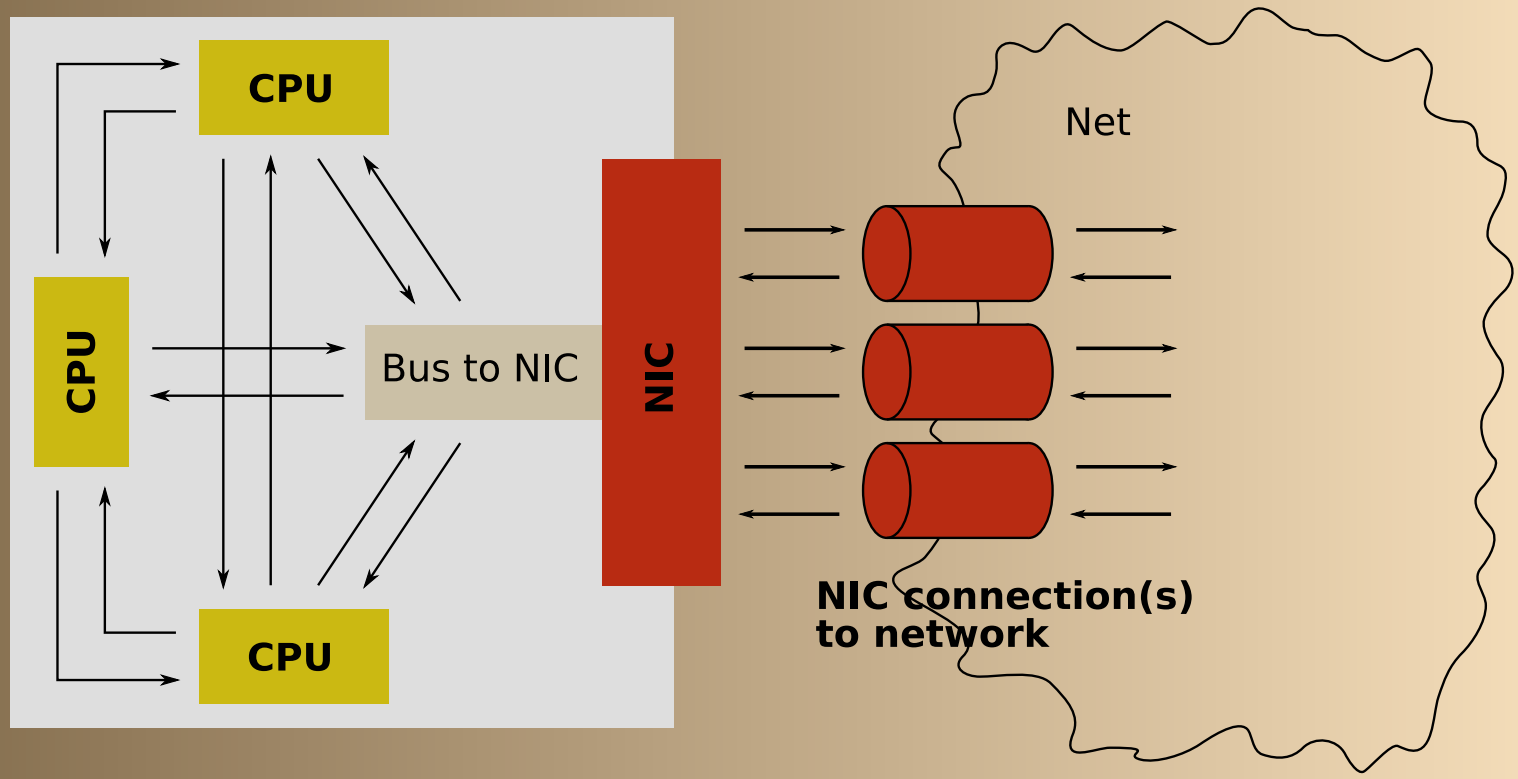
network



Improving the Model

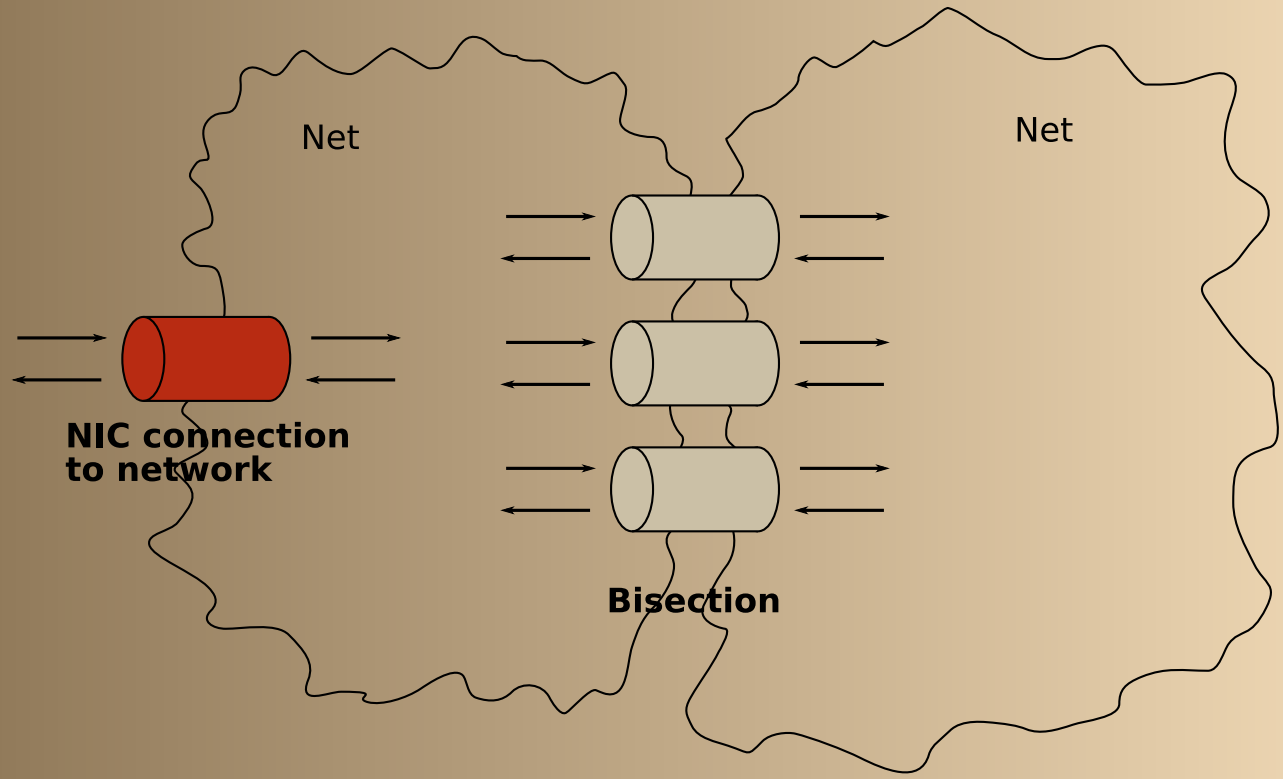
- Modeling message injection rate
- Multicore system
  - Shorter delays and less congestion between cores
- NIC throughput (when used by multiple cores)
- Modeling topological bisection bandwidth
  - Number of bisection links
- Number of connections between NIC and network
- Node allocation choice

Multicore System



- Contention for NIC access
- Cores do not have to be on same physical node

Bisection Bandwidth and Topology



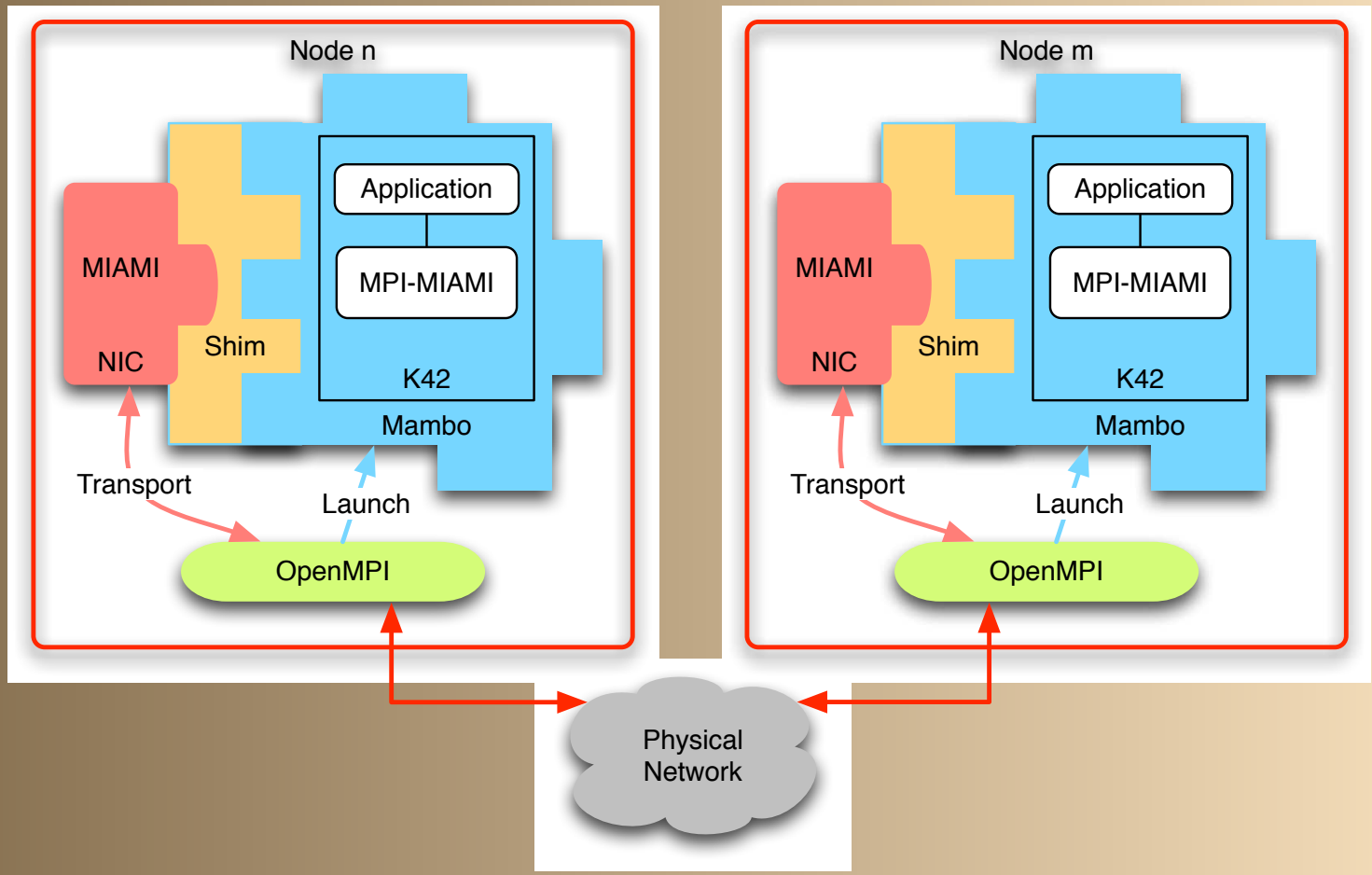
- Bisection is modeled through
  - Number of links, rank allocation algorithm, Link bandwidth and latency

Parameterization

- CPU speed factor: Time adjustment between MPI calls
- Number of bisection links
- Cross-over point to long MPI protocol
- MPI buffer space for short, core-to-core messages
- Bandwidth of an individual network link
- Latency of an individual network link
- Router fanout: Number of ports into network
- Bus bandwidth leading into the NIC
- Bus latency leading into the NIC
- Message interleaving on links
- Send overhead to another core
- Receive overhead from another core
- Send overhead through NIC
- Receive overhead from NIC
- Number of cores per NIC
- Core to NIC allocation: linear, round-robin, or random

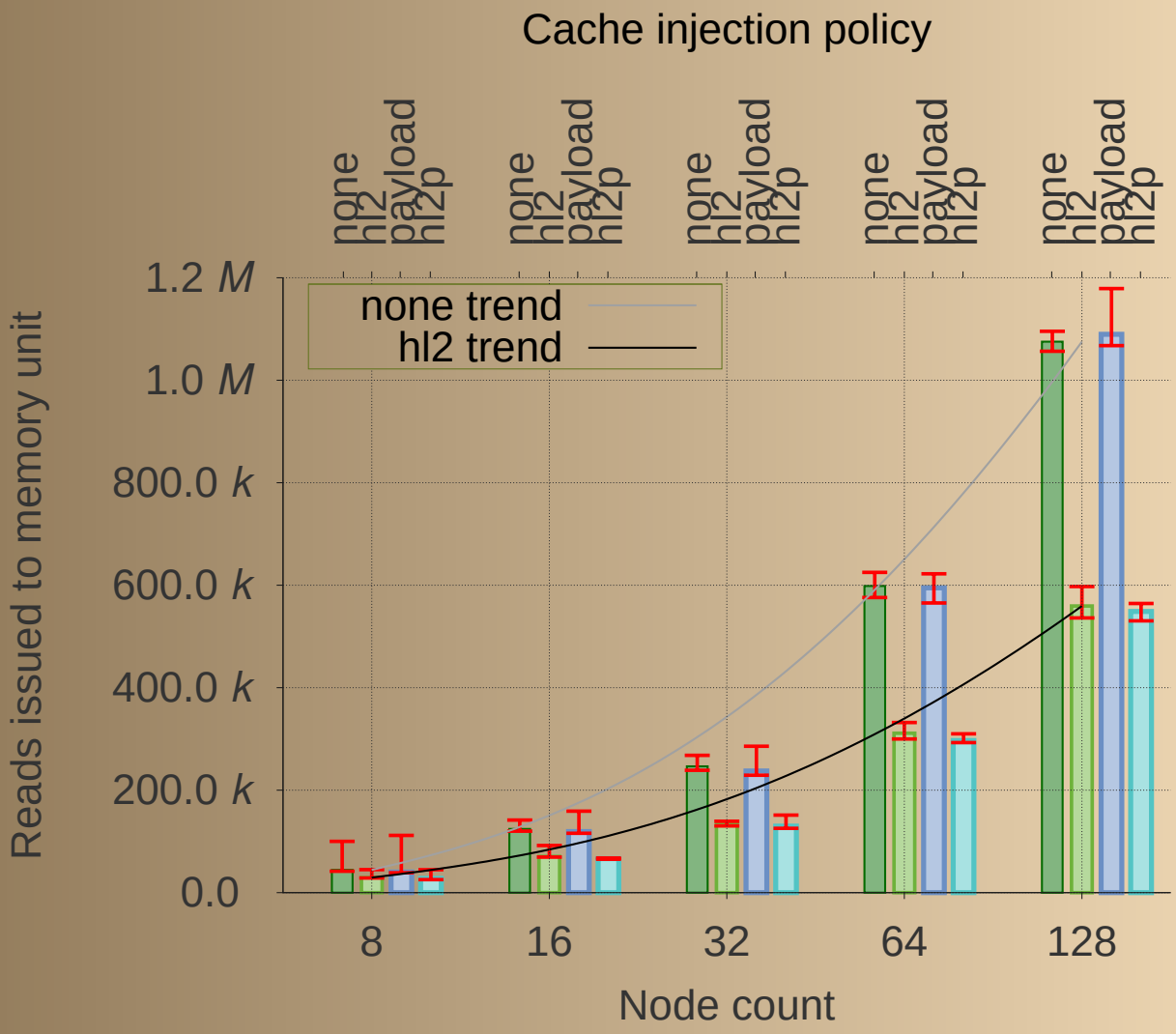
Cluster Simulator

Cycle-Accurate Simulation at Scale



Concepts

- Accurately simulate node (IBM's Mambo)
- Model network
- Ideal to study node characteristics at scale
- Synchronization interval: 50'000 cycles
  - Enough for accuracy, while still delivering good performance
- Fast-forward mode
  - No cache simulation until interesting part of application is reached
- Scalability
  - As good as application under simulation

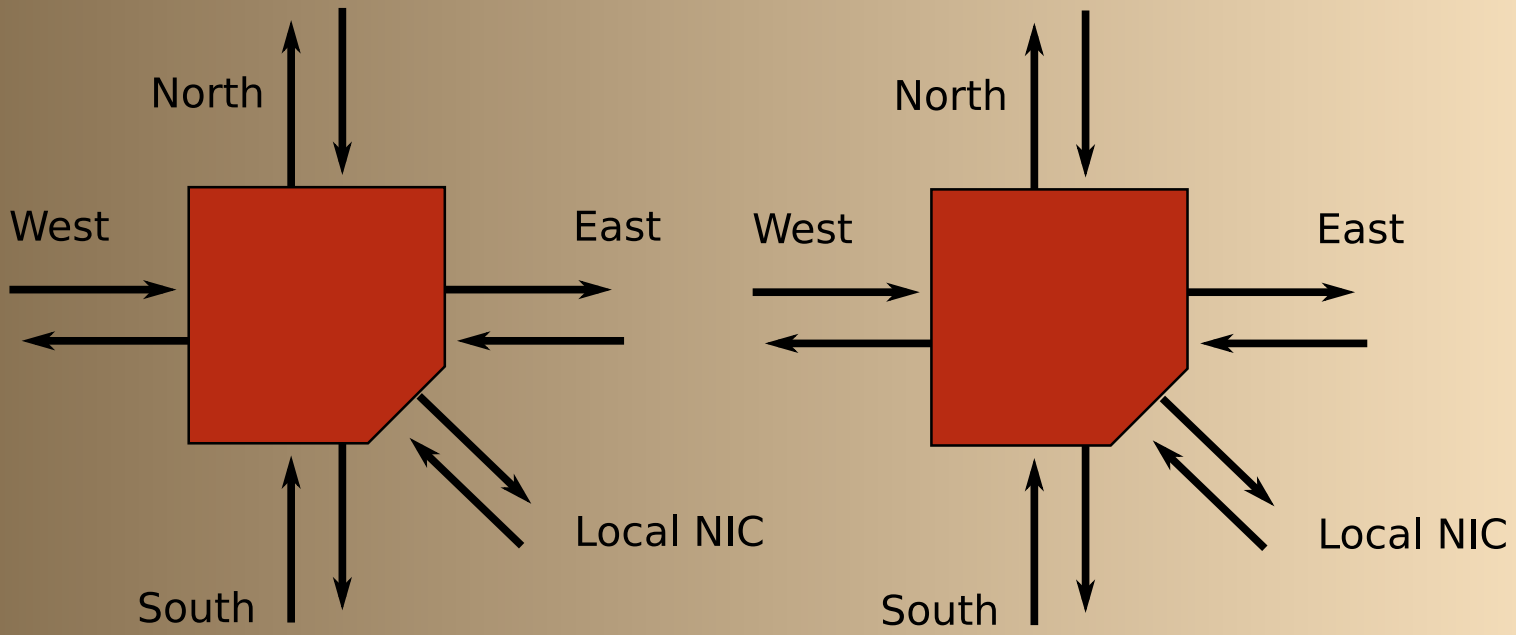


Cache injection lowers memory pressure and increases performance.

SST

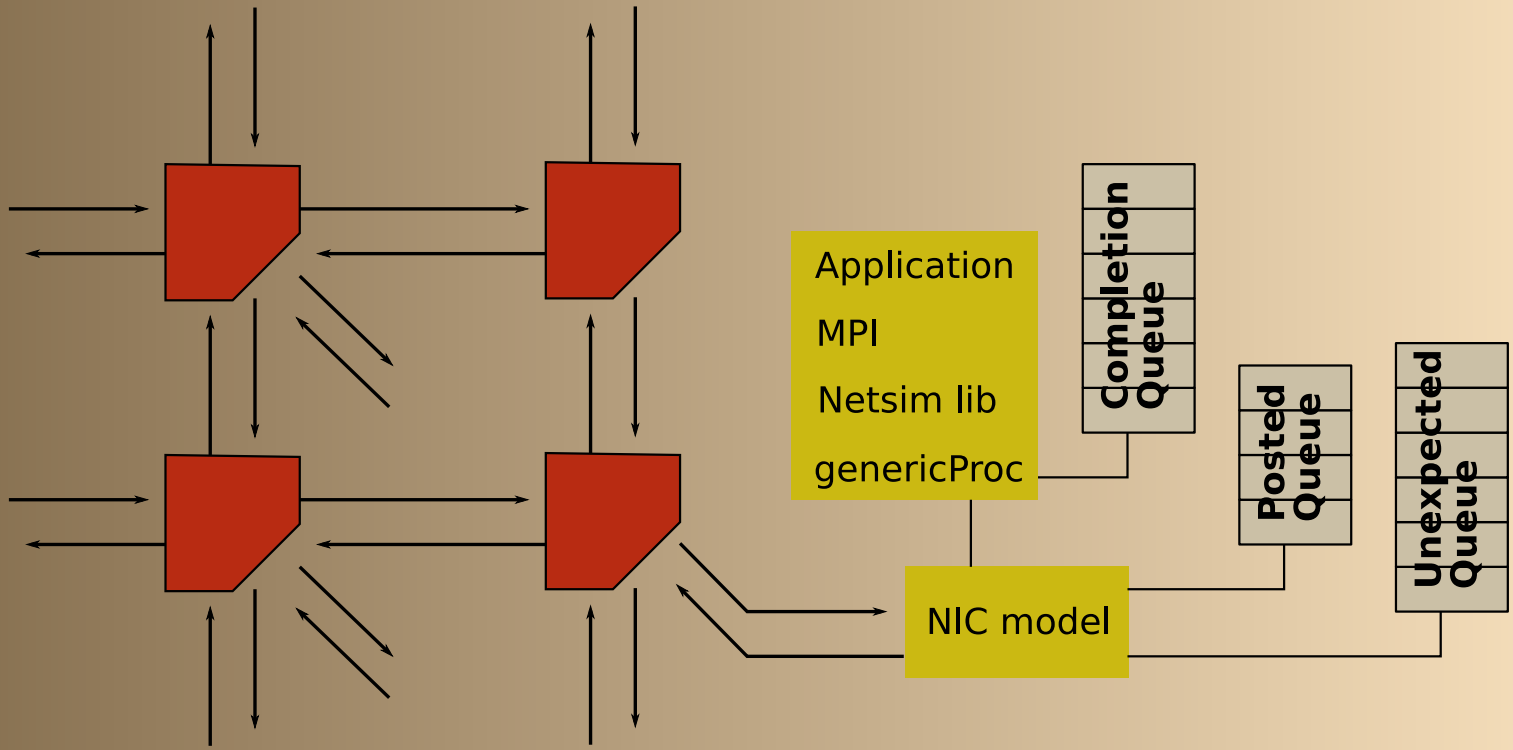
Network model in SST

- Apply Seshat and Cluster Sim lessons to SST
- Use genericProc component as a driver
  - Jeanine Cook's processor when available
  - Should work with any CPU component; e.g. trace-driven driver
- Router model
- Intelligent NIC



- Similar to ASCI Red router
- Source routing in X and Y direction
- Add hop delay
- Congestion not done yet, but should be easy
- Integrate with CA topology builder

Software Stack



- Three components
  - Router model, CPU simulator, NIC model
- Message matching in NIC model
- Delay model in NIC

The End

- Cluster Simulator:
  - Supercomputing'09 paper: *Instruction-Level Simulation of a Cluster at Scale*, Edgar Leon, Rolf Riesen, Arthur B. Maccabe, Patrick G. Bridges
- Seshat:
  - Cluster'06 paper: *A Hybrid MPI Simulator*, Rolf Riesen
  - CAC'06 paper: *Communication Patterns*, Rolf Riesen

Questions?